



**Daffodil**  
*International*  
**University**

## **Machine Learning Based Approach for Stroke Prediction**

**Supervised By:**

**Md. Rajib Mia**

Lecturer

Department of Software Engineering

Daffodil International University

**Submitted By:**

**Minara Afroze**

**ID: 181-35-309**

Department of SWE

Daffodil International University

This report has been submitted in fulfillment of the requirements for the Degree of B.Sc. in Software Engineering.

## THESIS DECLARATION

This is **Minara Afroze**, an undergraduate student from department of Software Engineering, Daffodil International University, Dhaka, Bangladesh. It hereby declares that this thesis titled “**Machine Learning Based Approach for Stroke Prediction**” has been done by me under the supervision of **Md. Rajib Mia, Lecturer**, Department of Software Engineering, Daffodil International University. It is also declared that neither this thesis nor any part of this has been submitted elsewhere for award of any degree.

Minara

**Minara Afroze**

ID: 181-35-309

Batch: 25th

Department of Software Engineering

Daffodil International University

**Supervised by:**



**Md. Rajib Mia**

**Lecturer,**

Department of Software Engineering,

Faculty of Science & Information Technology,

Daffodil International University.

## APPROVAL

This thesis titled on “Machine Learning Based Approach for Stroke Prediction”, submitted by **Minara Afroze (ID:181-35-309)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

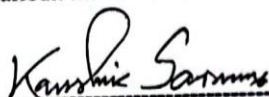
### BOARD OF EXAMINERS



**Chairman**

---

**Dr. Imran Mahmud**  
**Head and Associate Professor**  
 Department of Software Engineering  
 Faculty of Science and Information Technology  
 Daffodil International University



**Internal Examiner 1**

---

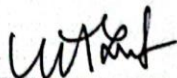
**Kaushik Sarker**  
**Associate Professor**  
 Department of Software Engineering  
 Faculty of Science and Information Technology  
 Daffodil International University



**Internal Examiner 2**

---

**Dr. Md. Fazla Elahe**  
**Assistant Professor**  
 Department of Software Engineering  
 Faculty of Science and Information Technology  
 Daffodil International University



**External Examiner**

---

**Mohammad Abu Yousuf, PhD.**  
**Professor**  
 Institute of Information Technology  
 Jahangirnagar University

## ACKNOWLEDGEMENT

First and foremost, I am grateful to Almighty Allah for giving me the ability to complete the final thesis. I'd like to thank my supervisor, Md. Rajib Mia, for his continuous help with my thesis and research work, which he provided through his motivation, energy, and sharing of information. His guidance supported me in finding research work solutions and obtaining my final theory. I want to express my heartfelt gratitude and respect to all of my teachers in the Software Engineering Department for their valuable guidance, kind help and support during the study. Last but not the least, I want to express gratitude to my family for always being there for me physically and spiritually throughout my life.

**Minara Afroze**

**ID: 181-35-309**

Department of Software Engineering,  
Faculty of Science & Information Technology,  
Daffodil International University.

## TABLE OF CONTENTS

<b>Thesis Declaration .....</b>	<b>i</b>
<b>Approval.....</b>	<b>ii</b>
<b>Acknowledgement.....</b>	<b>iii</b>
<b>List Of Tables.....</b>	<b>v</b>
<b>List Of Figures .....</b>	<b>v</b>
<b>Abstract .....</b>	<b>vi</b>
<b>Chapter 1: Introduction.....</b>	<b>1</b>
<b>Chapter 2: Literature Review .....</b>	<b>3</b>
<b>Chapter 3: Materials &amp; Methods.....</b>	<b>6</b>
<b>3.1 Data Description:.....</b>	<b>7</b>
<b>3.2 Data Preprocessing: .....</b>	<b>9</b>
<b>3.3 Supervised Machine Learning Algorithm: .....</b>	<b>10</b>
<b>3.4 Performance Evaluation Metrics:.....</b>	<b>14</b>
<b>Chapter 4: Experimental Results.....</b>	<b>16</b>
<b>Chapter 5: Conclusion .....</b>	<b>20</b>
<b>References.....</b>	<b>21</b>

## LIST OF TABLES

<b>TABLE 1: DATA DESCRIPTION.....</b>	<b>8</b>
<b>TABLE 2: BEST PARAMETER FOR ALL THE APPLIED ALGORITHMS .....</b>	<b>13</b>
<b>TABLE 3: PERFORMANCE METRICS &amp; FORMULA .....</b>	<b>15</b>
<b>TABLE 4: PERFORMANCE COMPARISON AMONG APPLIED CLASSIFIERS .....</b>	<b>16</b>

## LIST OF FIGURES

<b>FIGURE 1: WORKFLOW OF THE STUDY.....</b>	<b>6</b>
<b>FIGURE 2: PREDICTION METHODS OF KNN .....</b>	<b>12</b>
<b>FIGURE 3: ROC CURVE COMPARISON FOR ALL EMPLOYED ALGORITHMS. ....</b>	<b>17</b>
<b>FIGURE 4: PRECISION-RECALL CURVE COMPARISON FOR ALL ALGORITHMS USED...18</b>	

## ABSTRACT

When a blood vessel that supplies the brain with oxygen and nutrients becomes blocked by a clot or breaks, a stroke happens. Nowadays, it is the most common cause of death in the entire world. Every four minutes, someone become a victim of stroke and passes away, although 80% of stroke deaths may be avoided if we could recognize or anticipate them before they happened. Early stroke detection may be preferable to reducing the severity of the condition. Data science has played a significant role in the development of medical studies in recent years. To predict the chance of a stroke, several machine learning approaches are developed that use a patient's physical and physiological reporting data. The most significant risk factors for stroke in patients include age, cardiac disease, average blood sugar level, and hypertension. In this study, we employ Decision Tree, XG Boost, Light Gradient Boosting Machine (LGBM), Random Forest and K-nearest Neighbors learning as five machine learning algorithms to determine the most accurate model that can anticipate the risk of stroke and the dataset was collected through Kaggle. In this study, comparing with other machine learning algorithms utilized, the testing results indicate that the Random Forest algorithm has the maximum accuracy rate that is 96%.

**Keywords:** Ischemic stroke, Hemorrhagic stroke, Light Gradient Boosting Machine, Precision-Recall Curve · Random Forest · ROC Curve· Stroke Prediction.

## CHAPTER 1: INTRODUCTION

The demise of brain tissues occurs when oxygenation to certain areas of the brain is stopped or reduced because those tissues do not get the oxygenated blood they need. When blood supply to certain parts of the brain are interrupted, the tissues in those parts are deprived of oxygenation and nutrition. They are starting to expire due to running out oxygen. Stroke is the leading cause of death and disability globally. Stroke is a mental physical issue that occurs at any moment. To be successful, you must first recognize the issue and then successfully handle it. You also need to prevent further damage to the affected area of the brain as well as variety of other problems in various body parts. Ischemic and hemorrhagic strokes are the two forms that are recognized. While internal bleeding occurs when a weak blood artery explodes and bleeds into the nervous system, a biochemical stroke is hindered by clots. As per World Health Association (WHO) in consistently fifteen million people are suffering from stroke in worldwide and affected people are passing away every 4-5 minutes. As reported by the Centers for Disease Control and Prevention, stroke is the second most common reason of death globally and the sixth leading cause of death in the United States. About 11% of people around the world suffer from stroke, a nosocomial disease. A healthy lifestyle, avoiding smoking, and balancing normal BMI and sugar levels can all help to prevent stroke. Predicting stroke at the primary level can help to avoid permanent damage or death. This research work proposes an early prediction of stroke diseases by using different machine learning approaches. Machine learning approaches are being utilized to detect strokes in the early stages. Stroke can be avoided, if it can be anticipated sooner and necessary actions are taken. The goal of the study is to propose



an effective model based on that continuity that will allow both patients and doctors to predict stroke in its early stages using machine learning methods.

In this paper, we provide the following parts:

- Initially, we gathered open-source data to train and test each target model in order to determine which classification algorithm performed the best overall.
- After that, in order to increase the suitability of the ML models for better comprehension, we employed various preprocessing techniques.
- Next, in order to suggest more advanced technology to identify stroke early on, the highest performing classification methods are identified.
- Finally, after evaluating many classifiers, the optimum classification algorithm for stroke prediction is suggested in this paper.

The remaining sections of this study are organized as follows: The literature review is presented in chapter 2. The materials and methods used in this investigation are described in chapter 3 along with a description of the datasets. Chapter 4 contains a discussion of the tools and experimental outcomes. In the conclusion part, we represented conclusions and the future path of this study.

## CHAPTER 2: LITERATURE REVIEW

For the purpose of stroke prediction, many researchers have previously deployed machine learning-based methods. In a study to identify heart stroke diseases, Govindarajan et al. (P. Govindarajan, 2020) collected data from 507 patients and used a text mining and machine learning classifier combination. They employed Artificial Neural Networks, to train multiple machine learning algorithms for their study, and the SGD strategy provided the greatest value (95%).

The Cardiovascular Health Study (CHS) dataset was used by Sheetal et al. in their study (M. S. Singh and P. Choudhary, 2017) to compare several approaches with their methodology for stroke prediction. In order to develop a classification model in this paper, back propagation neural network classification technique is employed together with decision tree algorithm for feature selection and principal component analysis approach for dimension reduction. Their work was determined to provide an accuracy of 97.7% after analysis and comparison with other methods, although in order to reduce complexity, only those diseases that are not linearly spreadable are chosen. Furthermore, only 212 cases of stroke were found in 1824 samples, resulting in an imbalanced categorization.

Amini et al. (L. Amini, 2013) studied 50 risk variables for stroke, including diabetes, heart disease, smoking, hyperlipidemia, and alcohol consumption, and collected data from 807 physically active and unwell participants. The accuracy of the c4.5 decision tree approach was 95%, and the accuracy of K-nearest neighbor was 94%, these two techniques had the best efficiency.

A study on the calculation of the prediction of ischemic stroke was reported by Cheng et al. (C. A. Cheng, 2014). Two Artificial Neural Network models with accuracy of 79%

and 95% were employed in their study, which included 82 ischemic stroke patient health - related information.

A research study was conducted by Sung et al. (S.-F. Sung, 2015) to create a stroke intensity measure. They gathered information on 3577 patients who had recent ischemic strokes. They combined different data mining approaches with linear regression to create their predictive models. Its predictions exceeded the k-nearest neighbor method with an accuracy level of 95%.

A study was conducted by Cheon et al. (S. Cheon, 2019) to forecast stroke patient death. They employed 15099 individuals in their research to determine the frequency of strokes. To identify strokes, they applied a deep neural network technique. To obtain health record history and forecast stroke, the authors employed PCA. Their AUC value (area under the curve) is 83%.

Using machine learning, Monteiro et al. (M. Monteiro, 2018) conducted a research project to forecast the results of an ischemic stroke. They used this approach on a patient who died three months after admission in their study. They obtained an AUC value that was greater than 90%.

A study to identify an automated early ischemic stroke was conducted by Chin et al. (C.-L. Chin, 2017). The major objective of their study was to create a system utilizing CNN to automate primary ischemic stroke. To build and evaluate the CNN model, they gathered 256 pictures. They utilized data extension technique to elevate the obtained image in their system image preparation to remove the impossible stroke-causing area. 90% accuracy has been reported for their CNN approach.

A study was conducted by Adam et al. (S. Y. Adam, 2016) to categorize ischemic stroke. To classify ischemic stroke, they utilized two different approaches: a k-nearest

neighbor approach and a decision tree algorithm. They found that when used to categorize strokes by medical professionals, the decision tree approach was more practical. With a 90% accuracy rate, the decision tree method classified strokes more accurately.

Based on the preceding explanation, it is apparent that more advanced technology is required to overcome existing restrictions and create more accurate and upgraded technology. In this regard, the study seeks to identify the highest performing classifier in order to develop a more advanced model to predict stroke in its early stages.

## CHAPTER 3: MATERIALS & METHODS

As already stated, it is obvious that there is still an opportunity to create an efficient machine learning-based model that can identify stroke in its initial phases. We tested different machine learning models using an openly accessible dataset in this study. In previous research, the majority of authors used a significant method to identify stroke disease. Nonetheless, we used five alternative approaches to compare the findings to earlier studies. The results of this study and analyses are summarized in the following section.

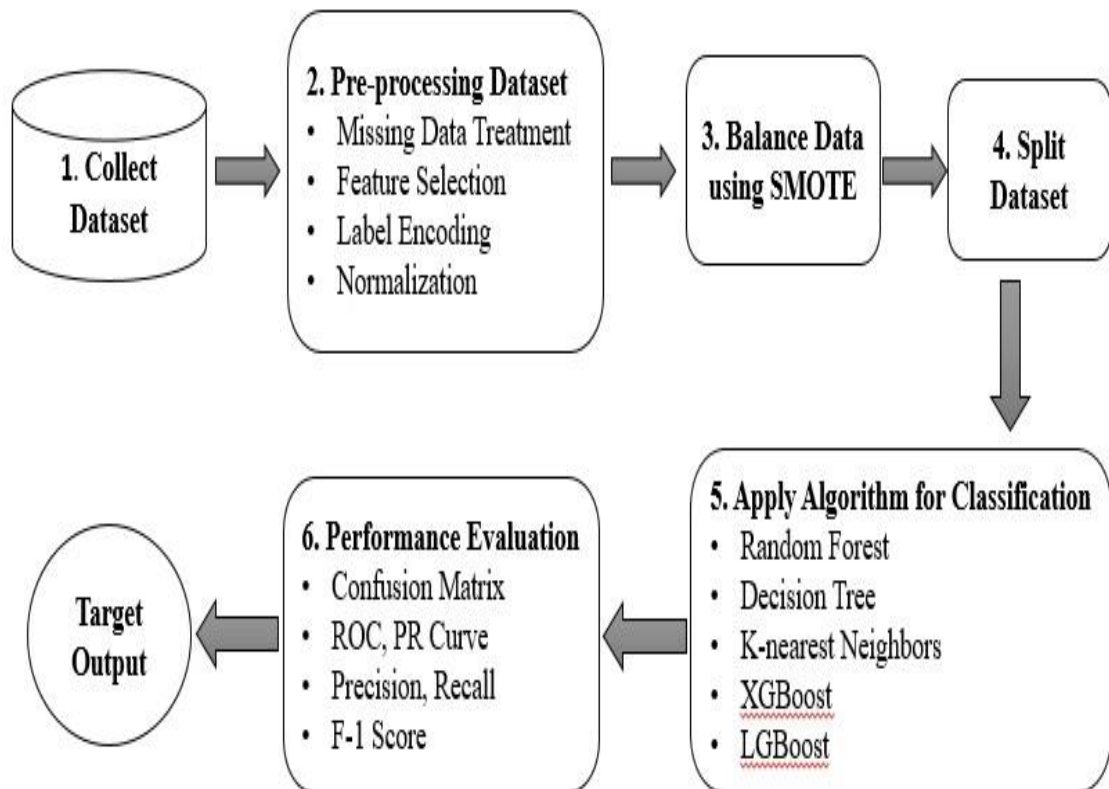


Figure 1: Workflow of the study

The research process and workflow for developing an effective machine learning based system that can detect stroke for a patient with more accuracy are outlined in figure 1. The suggested model can be used to develop more sophisticated systems that will be able to predict stroke depending on certain detail information or characteristics of a patient.

### **3.1 Data Description:**

To construct and establish our model, we utilized a Kaggle dataset (kaggle). The dataset consists of 12 characteristics with a total of 5110 cases, of which 2994 are female patients, 2115 are male patients, and 1 instance was unspecified. Approximately half of the patients in this dataset are located in cities, while the rest are located in rural areas. There are 249 persons with stroke and the others have no probability of having a stroke. 276 of the patients had a history of cardiac disease, and 17% of the patients had experienced a stroke. Despite the fact that this dataset is not balanced, it generates a favorable output. Table 1 contains required data on the dataset as well as feature information.

Table 1: Data Description

Attribute	Data Type	Description
ID	Numerical	There is a distinct ID for each patient.
Gender	Categorical	Identifies the gender of patients
Age	Numerical	Describe the age of the patient.
Work Type	Categorical	Informs about the working category
Ever married	Binary	Mention about marital status.
Residence Type	Categorical	indicate the living environment
Heart Disease	Binary	Whether or not the patients have cardiac disease.
Hypertension	Binary	The presence or absence of hypertension in the patients
BMI	Numerical	Represent the patient's Body mass index (BMI) value
Glucose Level	Numerical	It indicates the average level of glucose.
Stroke	Binary	Informs about the prior stroke
Smoking Status	Categorical	Gives patients smoking scenario

## 3.2 Data Preprocessing:

For machine learning to generate better outcomes, the raw data is often unreliable and ineffective. Data preprocessing is therefore a vital operation that must be carried out to prepare data for better analytical results in order to improve training and testing performance. Before developing a model, data preprocessing is essential to eliminate unnecessary noise and outliers from the dataset, which can cause a deviation from good training. This stage addresses everything that prevents the model from performing more efficiently.

**Missing Value Treatment:** Initially, mean and mode statistical approaches are employed in the preprocessing stage to deal with missing variables. If there are any null values in the dataset, they are filled. The mean of the column data is used to fill the null values in the column labeled "BMI" in this instance.

**Label Encoding:** For categorical features, label encoding is executed. Label encoding converts categorical values in a dataset into integer values that the machine can interpret. Strings need to be encoded to integers since machines are always operated on numbers. The gathered dataset contains five columns with categorical data. All of the categorical data are encoded when label encoding is executed, and the entire dataset becomes numerical data.

**Handling Imbalanced Data:** The selected dataset for predicting strokes is highly imbalanced. 249 people experienced strokes out of 5110 cases, whereas the remaining patients were at no risk. It shows that the dataset is unbalanced. Without proper handling the imbalanced data, results will be inaccurate and predictions will be ineffective. Therefore, handling this imbalanced data is necessary before creating an effective



model. Since the dataset was unbalanced, the dataset was then balanced using the Synthetic Minority Oversampling Technique (SMOTE).

**Splitting the Data:** In order to split a dataset, it must be divided into training and testing categories. The split approach is utilized for training and testing in this paper.

### **3.3 Supervised Machine Learning Algorithm:**

#### **Random Forest:**

Random Forest is a type of supervised learning approach that can be used to handle regression and classification problems. The development of the RF classification method is mostly based on several decision trees. To construct each tree, Random Forest randomly chooses features and applies the bagging and booting technique. The accuracy of this uncorrelated forest is higher than that of a single tree. It generates decision trees from various samples by using the majority vote for categorization and the mean for prediction. It produces superior outcomes for classification issues. One of the most interesting characteristics of random forest is its flexibility. Random forests provide the highest reliability of any classification method currently available, and it may be utilized in both regression and classification applications. The random forest technique is also able to deal with huge datasets that contain hundreds of parameters (M. S. Singh and P. Choudhary, 2017). Random forest is absolutely essential for preventing overfitting problems.

**XG Boost:**

The gradient-enhanced decision tree variant known as XG Boost is an open-source and free version that is strengthened and more stable. Although it is a classification technique of the ensemble variety, it can also be utilized for tasks involving regression. It has parallelization capabilities (Ramraj, 2016). Because of its tremendous adaptability and versatility, XG Boost dominates structured or tabular datasets for classification, regression, and predictive modeling. It has automatic missing value handling, a block structure to enable concurrent tree building, and then a continuous training method to allow for extra boosting of previously fitted models on new data.

**K-Nearest Neighbors (KNN):**

One type of supervised machine learning is the KNN classifier, which uses a straightforward and effective identification approach to deal with classification and regression issues. Since it is non-parametric, pattern recognition is its main application (Suguna, 2010). The already-existing labelled data is used to train KNN. Applications requiring deep domain knowledge are best suited to the KNN technique. K denotes the number of instances that can be regarded a raw data point's neighbor. Figure 2 illustrates the prediction process of KNN.

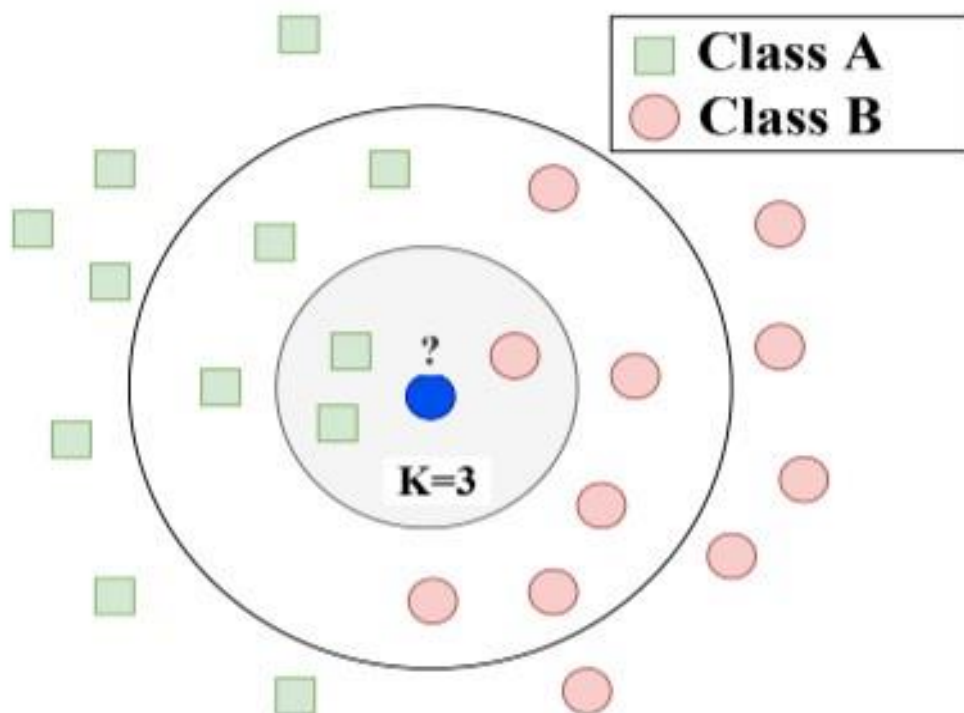


Figure 2: Prediction methods of KNN

### **Light Gradient Boosting Machine:**

One type of ensemble learning supervised classification method is the Light Gradient Boosting Machine (LGBM). LGBM relies on decision tree algorithms, which are a quick, scalable, and incredibly efficient framework for gradient boosting. When exceptional precision is required, it focuses on correctness of outcomes and can manage massive volumes of data (Ahamed, 2021). Gradient-based one-side sampling and exclusive feature bundling are the two types of methods used by LGBM. Consequently, these two models help the system function properly.

### Best Parameter for All the Applied Algorithms:

The parameters for all the algorithms presented in Table 2 include those that were hyper-tuned for all the applicable parameters. The applicable algorithms in this study all work effectively in these conditions.

Table 2: Best Parameter for All the Applied Algorithms

Classifier	Best Parameter
Random Forest	'max depth': 100, 'max features': 'auto', 'min samples leaf': 1, 'n estimators': 200 'learning rate': 0.1, 'max depth': 10,
LGBM classifier	'min child samples': 5, 'num iterations': 300, 'num leaves': 80, 'reg alpha': 0 'n estimators': 300, 'colsample bytree': 0.5,
XGB classifier	'gamma': 0.25, 'learning rate': 0.1, 'max depth': 7, 'reg lambda': 10, 'scale pos weight': 5, 'subsample': 0.8
Decision Tree	'criterion': 'gini', 'max depth': 10, 'min samples leaf': 50
K-Nearest Neighbors	'leaf size': 3, 'n neighbors': 13, 'p': 1

### 3.4 Performance Evaluation Metrics:

All applicable classification methods are assessed employing four components: accuracy, precision, recall, and f1-score. The proportion of successfully classified events to total projected occurrences is specified as accuracy. One of the easiest Classification metrics to apply is the accuracy metric. The percentage of accurate positive class forecasts is known as precision. The accuracy metric's drawback is resolved by employing the precision metric. The recall is the number of correct positive class predictions made out of all correct positive examples in the dataset. The accuracy and recall standards are combined into a single metric to generate the F1 score. The F1 score was also created to function well with unbalanced data. In these performance metrics, the following formula is employed to evaluate systems (Rahaman, 2022).

Below, TP and TN stand for true positives and true negatives, respectively. Similarly, false positives and false negatives are indicated by the acronyms FP and FN, accordingly.

Table 3: Performance Metrics &amp; Formula

<b>Metric</b>	<b>Formula</b>
Accuracy	$(TP+TN) / (TP+TN+FP+FN)$
Precision	$TP / (TP+FP)$
Recall	$TP / (TP+FN)$
F1 score	$(2*TP) / (2*TP+FN+FP)$

Here,

TP = True Positive, Actual result is true and obtained result is also true.

FP = False Positive, Actual result is false and obtained result is also false.

TN = True Negative, Actual result is true but obtained result is false.

FN = False Negative, Actual result is false but obtained result is true.

## CHAPTER 4: EXPERIMENTAL RESULTS

In this work, supervised machine learning approaches were employed to analyze the data using Python (version 3.8.5). In this study, five supervised machine classifiers decision tree, random forest, XG Boost, LGBM and K-Nearest Neighbor were used. The best fit classifiers were chosen after analyzing their effectiveness.

All of the implemented methods performance outcomes are shown in Table 4, which is given below.

Table 4: Performance Comparison among applied classifiers

Algorithm	Accuracy	Precision	Recall	F1-score
Random Forest	95.84	0.96	0.96	0.96
LGBM	95.63	0.96	0.96	0.96
Decision Tree	93.68	0.94	0.94	0.94
K-nearest Neighbors	91.47	0.92	0.92	0.92
XG Boost	89.61	0.90	0.90	0.90

According to the table, Random Forest and LGBM showed the best performance, with accuracy rates of 95.84% and 95.63%, respectively, and 0.96 precision, recall, and f1 scores for both Random Forest and LGBM, where XG Boost performs worse than all other classifiers. On the other hand, Decision Tree and K-nearest Neighbors showed the accuracy rate 93.68% and 91.47% respectively. So, It is strongly suggested to use Random Forest for stroke prediction because it has higher accuracy than LGBM.

The ROC curve for each of the used classification techniques is shown in Figure 3, which is given below.

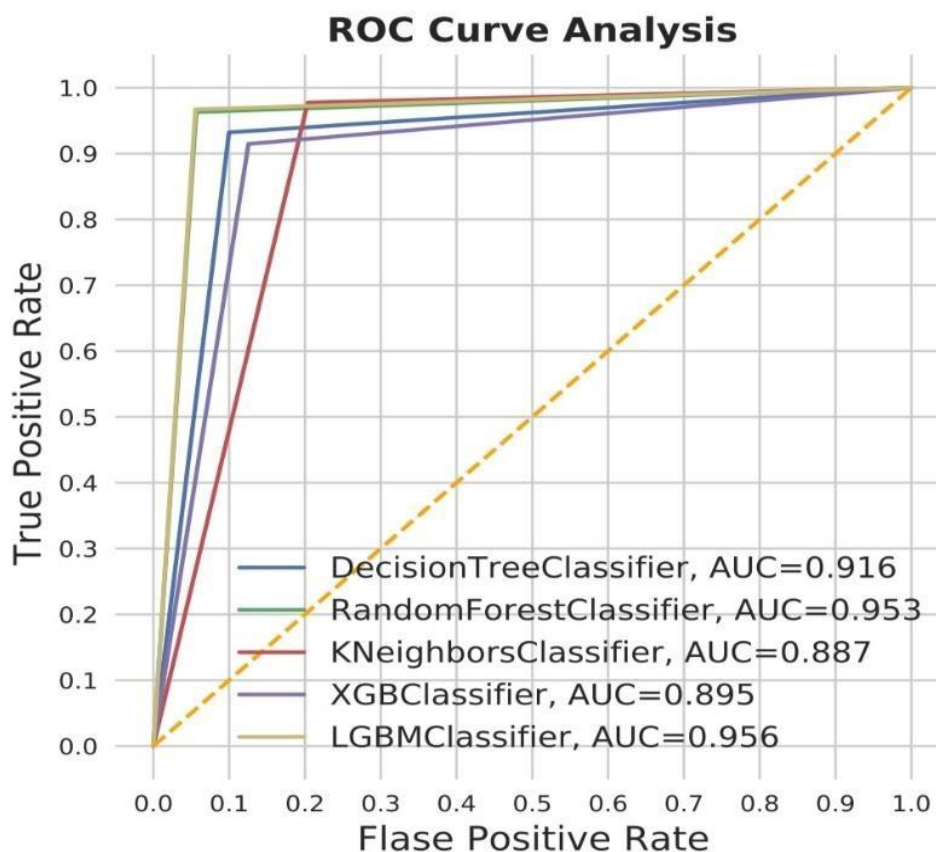


Figure 3: ROC curve comparison for all employed algorithms.



From the figure .3 we can show that the area under the ROC curve that is covered by K-Nearest Neighbor is the lowest, while LGBM, which has values of 0.956, covers the largest area under the ROC curve. Additionally, Random Forest generates a result that is 0.953 points nearer to LGBM.

The PR curve for each of the used classification algorithms is shown in Figure 4, which is given below.

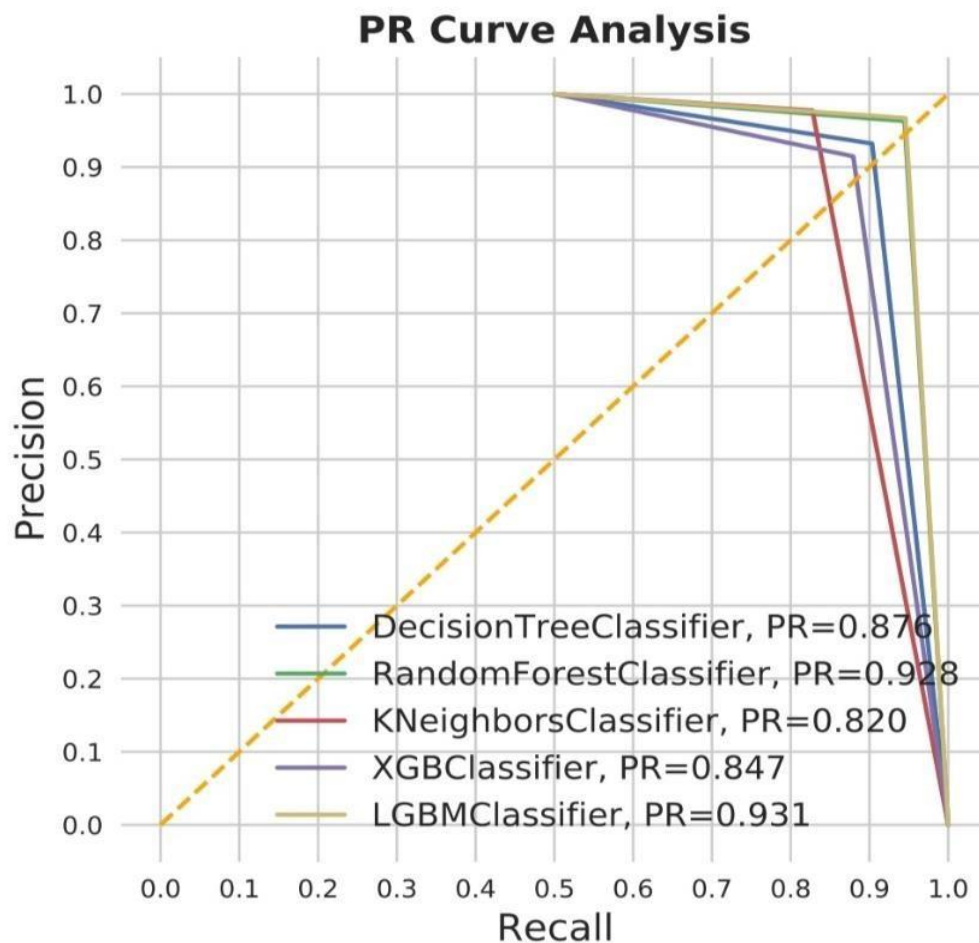


Figure 4: Precision-Recall curve comparison for all algorithms used

Form the figure .3 we can show that in terms of both area under the PR curve, LGBM covers the largest ground with a value of 0.931, while K-Nearest Neighbor has the least area covered with a value of 0.820. Decision Tree classifier offer a result that is 0.876 points and XGB classifier offers a result that is 0.847. Additionally, Random Forest offers a result that is 0.928 points closer to LGBM. Both Random Forest and LGBM showed satisfactory performance in this case.

## CHAPTER 5: CONCLUSION

Stroke is a serious medical condition that needs to be handled immediately to prevent future complications. The construction of a powerful machine learning-based system that can aid in the early identification and prevention of stroke's disastrous effects. Doctors will be able to decide the treatment instantly by detecting stroke disease using the proposed machine learning model. Using a variety of machine learning techniques, this study investigated how to predict strokes with accuracy using a range of important variables. With classifier performance levels of 95.84 percent for random forest and 95.64 percent for LGBM, the classification algorithm beats the other approaches analyzed. The study showed that when cross-validation measures are used to predict brain strokes, the Random forest and LGBM techniques outperform other approaches. By utilizing a larger population and machine learning techniques like Ada Boost, SVM, and Bagging, the framework models may eventually be expanded. Both the reliability and aesthetic appeal of the software are enhanced by doing this. Doctors and clinicians would be able to anticipate strokes and treat patients appropriately with the suggested strategy.

## REFERENCES

1. <https://www.kaggle.com/datasets/fedesoriano/stroke-prediction-dataset>
2. Ahamed, B. S., & Arya, S. (2021). LGBM classifier based technique for predicting type-2 diabetes. *European Journal of Molecular & Clinical Medicine*, 8(3), 454-467.
3. Cheng, C. A., Lin, Y. C., & Chiu, H. W. (2014, January). Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks. In *ICIMTH* (pp. 115-118).
4. Chin, C. L., Lin, B. J., Wu, G. R., Weng, T. C., Yang, C. S., Su, R. C., & Pan, Y. J. (2017, November). An automated early ischemic stroke detection system using CNN deep learning algorithm. In *2017 IEEE 8th International conference on awareness science and technology (iCAST)* (pp. 368-372). IEEE.
5. Amini, L., Azarpazhouh, R., Farzadfar, M. T., Mousavi, S. A., Jazaieri, F., Khorvash, F., ... & Toghianfar, N. (2013). Prediction and control of stroke by data mining. *International journal of preventive medicine*, 4(Suppl 2), S245.
6. Monteiro, M., Fonseca, A. C., Freitas, A. T., e Melo, T. P., Francisco, A. P., Ferro, J. M., & Oliveira, A. L. (2018). Using machine learning to improve the prediction of functional outcome in ischemic stroke patients. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(6), 1953-1959.
7. Singh, M. S., & Choudhary, P. (2017, August). Stroke prediction using artificial intelligence. In *2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON)* (pp. 158-161). IEEE.

8. Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3), 817828.
9. Rahaman, M. A., Ali, M. M., Ahmed, K., Bui, F. M., & Mahmud, S. H. (2022, March). Performance Analysis between YOLOv5s and YOLOv5m Model to Detect and Count Blood Cells: Deep Learning Approach. In *Proceedings of the 2nd International Conference on Computing Advancements* (pp. 316-322).
10. Ramraj, S., Uzir, N., Sunil, R., & Banerjee, S. (2016). Experimenting XGBoost algorithm for prediction and classification of different datasets. *International Journal of Control Theory and Applications*, 9(40).
11. Cheon, S., Kim, J., & Lim, J. (2019). The use of deep learning to predict stroke patient mortality. *International journal of environmental research and public health*, 16(11), 1876.
12. Adam, S. Y., Yousif, A., & Bashir, M. B. (2016). Classification of ischemic stroke using machine learning algorithms. *International Journal of Computer Applications*, 149(10), 26-31.
13. Sung, S. F., Hsieh, C. Y., Yang, Y. H. K., Lin, H. J., Chen, C. H., Chen, Y. W., & Hu, Y. H. (2015). Developing a stroke severity index based on administrative data was feasible using data mining techniques. *Journal of clinical epidemiology*, 68(11), 1292-1300.
14. Suguna, N., & Thanushkodi, K. (2010). An improved k-nearest neighbor classification using genetic algorithm. *International Journal of Computer Science Issues*, 7(2), 18-21.

## Turnitin Originality Report

Processed on: 10-Nov-2022 16:06 +06  
 ID: 1950076443  
 Word Count: 4383  
 Submitted: 1

181-35-309 By Minara Afroze

Similarity Index

23%

Similarity by Source

Internet Sources: 16%  
 Publications: 12%  
 Student Papers: 11%

2% match (Internet from 25-Oct-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5684/171-35-1831%20%2817%20%29.pdf?isAllowed=y&sequence=1>

2% match (Internet from 15-Nov-2021)

[https://www.researchgate.net/publication/348133587\\_Performance\\_Analysis\\_of\\_Machine\\_Learning\\_Approaches\\_in\\_Stroke\\_Prediction](https://www.researchgate.net/publication/348133587_Performance_Analysis_of_Machine_Learning_Approaches_in_Stroke_Prediction)

2% match (Internet from 14-Sep-2022)

<https://www.hindawi.com/journals/jhe/2021/7633381/>

1% match (Internet from 26-Oct-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8573/181-35-348.pdf?isAllowed=y&sequence=1>

1% match (Internet from 26-Oct-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8565/181-35-289.pdf?isAllowed=y&sequence=1>

1% match (student papers from 20-Aug-2022)

[Submitted to The Robert Gordon University on 2022-08-20](#)

1% match (student papers from 20-May-2022)

[Submitted to University of Teesside on 2022-05-20](#)

1% match (Internet from 17-Dec-2021)

<https://www.coursehero.com/file/101339879/P12751-11pdf/>

1% match (Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, Mohammad Monirujjaman Khan. "Stroke Disease Detection and Prediction Using Robust Learning Approaches", Journal of Healthcare Engineering, 2021)  
[Tahia Tazin, Md Nur Alam, Nahian Nakiba Dola, Mohammad Sajibul Bari, Sami Bourouis, Mohammad Monirujjaman Khan. "Stroke Disease Detection and Prediction Using Robust Learning Approaches", Journal of Healthcare Engineering, 2021](#)

1% match (M. Sheetal Singh, Prakash Choudhary. "Stroke prediction using artificial intelligence", 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference (IEMECON), 2017)

[M. Sheetal Singh, Prakash Choudhary. "Stroke prediction using artificial intelligence", 2017 8th Annual Industrial Automation and Electromechanical Engineering Conference \(IEMECON\), 2017](#)

1% match ("ICT Systems and Sustainability", Springer Science and Business Media LLC, 2023)  
["ICT Systems and Sustainability", Springer Science and Business Media LLC, 2023](#)

1% match (student papers from 05-Dec-2021)

[Submitted to Coventry University on 2021-12-05](#)

1% match (Internet from 02-Jun-2022)

[https://mdpi-res.com/d\\_attachment/entropy/entropy-23-00763/article\\_deploy/entropy-23-00763.pdf](https://mdpi-res.com/d_attachment/entropy/entropy-23-00763/article_deploy/entropy-23-00763.pdf)

1% match (student papers from 03-Aug-2022)

[Submitted to University of Cape Town on 2022-08-03](#)

< 1% match (Internet from 15-Aug-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8402/181-35-2332.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 12-Jun-2022)

<http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/8152/171-35-198%20%2813%25%29.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 30-Aug-2022)

[https://www.researchgate.net/publication/359649144\\_Analisis\\_Perbandingan\\_Algoritma\\_C45\\_Dan\\_Naive\\_Bayes\\_Dalam\\_Memprediksi\\_Penyakit](https://www.researchgate.net/publication/359649144_Analisis_Perbandingan_Algoritma_C45_Dan_Naive_Bayes_Dalam_Memprediksi_Penyakit)

< 1% match (student papers from 20-May-2022)

[Submitted to University of Teesside on 2022-05-20](#)

< 1% match (student papers from 30-Aug-2022)

[Submitted to Liverpool John Moores University on 2022-08-30](#)

< 1% match (student papers from 31-Aug-2022)

[Submitted to Liverpool John Moores University on 2022-08-31](#)

< 1% match (Atul Kumar Uttam. "Analysis of Uneven Stroke Prediction Dataset using Machine Learning", 2022 6th International Conference on Intelligent Computing and Control Systems (ICICCS), 2022)

[Atul Kumar Uttam. "Analysis of Uneven Stroke Prediction Dataset using Machine Learning", 2022 6th International Conference on Intelligent Computing and Control Systems \(ICICCS\), 2022](#)

